Estimated Covariance Matrices Associated with Calibration

Luca Sartore* Kelly Toppin† Clifford Spiegelman‡

Abstract

Surveys often provide numerous estimates of population parameters. Some of the population values may be known to lie within a small range of values with a high level of certainty. Calibration is used to adjust survey weights associated with the observations within a data set. This process ensures that the "sample" estimates for the target population totals (benchmarks) lie within the anticipated ranges of those population values. The additional uncertainty due to the calibration process needs to be captured. In this paper, some methods for estimating the variance of the population totals are proposed for an algorithmic calibration process based on minimizing the L_1-norm relative error. The estimated covariance matrices for the calibration totals are produced either by linear approximations or bootstrap techniques. Specific data structures are required to allow for the computation of massively large covariance matrices. In particular, the implementation of the proposed algorithms exploits sparse matrices to reduce the computational burden and memory usage. The computational efficiency is shown by a simulation study.

Key Words: Variance, Estimation, Calibration, Weighting, Survey, Census

1. Introduction

The idea of calibration was initially introduced by Lemel (1976) and further developed by Deville and Särndal (1992) to improve the representativeness of a surveyed sample. This improvement is attained by providing an optimal weight for each sample unit. Calibration weighting techniques adjust probability-sampling weights in order to improve the precision of the population estimate by taking into account sampling frame errors and nonresponse. The optimal weights are computed to be as close as possible to the original sampling design weights by forcing the final estimate to be consistent with auxiliary measurements. This is accomplished by solving a system of linear equations, which is formulated as follow:

$$y = Aw, \tag{1}$$

where y denotes the vector of n point targets (or benchmarks), A represents the matrix $n \times p$ of collected data, and w is the vector of p unknown calibration weights to compute. Traditional calibration algorithms produce non-integer weights. At the National Agricultural Statistics Service, we do calibration with integer weights, which are estimated by solving an integer programming problem (Sartore and Toppin, 2016) such that

$$w \in \mathcal{N} \subset \mathbb{N}^p, \tag{2}$$

where the subset $\mathcal{N}$ denotes the set of integer weights which satisfy given constraints.

The solution of the linear system in (1) is based on the optimization of a penalized arbitrary differentiable loss function $L(\cdot)$ such that the constraints in (2) are satisfied, i.e.

$$\hat{w} = \arg \min_{w \in \mathcal{N} \subset \mathbb{N}^p} L(y - Aw) + \lambda P(w), \tag{3}$$

*National Institute of Statistical Sciences, 19 T.W. Alexander Drive, P.O. Box 14006, Research Triangle Park, NC 27709-4006, lsartore@niss.org

†National Agricultural Statistics Service, United States Department of Agriculture, 1400 Independence Ave. SW, Washington, DC 20250

‡Texas A&M University, 3135 TAMU, College Station, TX 77843-3135

where λ is a positive factor, and $P(\cdot)$ is a penalty that forces the optimal solution to be closed to the initial weights, which are provided by a dual-system estimation methodology. The optimal weights are then obtained through a stepwise procedure, which exploits the gradient of the loss function $L(\cdot)$.

In order to achieve robust results, the loss function $L(\cdot)$ can be formulated by considering variations of the L_1-norm, such as the summations of absolute errors or relative errors. These objective functions lead to non-smooth calibration estimators, and therefore, the jackknife variance estimators discussed by Wolter (1985) and some other re-sampling techniques are inconsistent (Furno, 1998). Several methods were proposed in the literature when the loss function is formulated as a summation of square residuals: e.g. the linearized variance estimation was initially introduced by Fuller (1975) for a simple regression estimator. Särndal et al. (1989) developed a technique for the variance estimation of a general regression estimator. Kott (1999) discussed a delete-a-group jackknife approach for providing unbiased estimates within a stratified survey sample. Singh and Folsom (2000) proposed a bias correction of the variance based on post-stratification adjustments. Valliant (2007) compared linearization and replication methods, and mentioned the effect of item imputation on the estimated variance. A comprehensive review about variance estimators for calibration and their computational complexities are described in Kott (2016).

In this article, classical estimation approaches are adjusted to estimate high-dimensional covariance matrices associated with calibration weights and totals. In section 2, a linear approximation technique is described in order to allow for the inversion of a high-dimensional sparse covariance matrix. Two resampling methods are then considered in section 3. In particular, the parametric bootstrap is presented in section 3.1 to provide consistent estimates with the properties of a covariance matrix. The non-parametric bootstrap is explained in section 3.2 by allowing for a more efficient computational strategy. A simulation study and its results are discussed in section 4, where the proposed methods are compared. The achievements, technical issues, assumptions and future research will be addressed in section 5. Finally, the conclusions are presented in section 6.

2. Linear approximation

Under the assumption of a meaningful correspondence between the loss function and a probabilistic model for the errors $y - Aw$, one can justify an interpretation of $L(\cdot)$ as a negative log-likelihood. Thus the covariance matrix can be calculated through the inversion of the information matrix. Such a matrix corresponds to the Hessian H of the objective function used in (3), and it is expected to be positive definite when the optimal weights are a local minimum. The variance of the calibration weights is given by $\mathsf{VAR}(\hat{w}) = H^{-1}$, where H^{-1} denotes the inverse matrix of H. This approach suffers from two computational problems: the first is related to the rank of the Hessian matrix, while the second is due to computational capabilities of storing the results of a very high-dimensional matrix (in particular when $p \gg n$).

The first problem arises when the Hessian is not a full-rank matrix. To address this problem, the following approximation of the Hessian is proposed:

$$H \approx -A^\top J(w) + \lambda I_p,$$

where I_p is the identity matrix of size $p \times p$. This term is used to allow for the inversion of the Hessian. Furthermore, $J(w)$ denotes an $n \times p$ sparse matrix-valued function, whose resulting rows are calculated by the first partial derivatives of the gradient $\nabla_{y-Aw} L(\cdot)$ with respect to the calibration weights. Moreover, another approximation is required when $L(\cdot)$ is not twice differentiable, and when this is the case, a generic entry of the matrix-valued

function $J(w)$ can be computed by the use of numerical derivatives, or by providing a smooth approximation of the gradient $\nabla_{y-Aw}L(\cdot)$.

Once all the terms to compute the Hessian are available, a large matrix must be inverted. Although the Hessian is not dense, the usual algorithms for inverting sparse matrices require extraordinary computational capabilities. However, the following approximation is based on the Woodbury's matrix equality (1950) investigated successively by Henderson and Searle (1980), and it can be exploited to obtain H^{-1} in a reasonable time with efficient use of memory:

$$\begin{aligned}\left(\lambda I_p - A^\top J(w)\right)^{-1} &= \frac{1}{\lambda} I_p \quad + \quad \frac{1}{\lambda} A^\top \left(\lambda I_n - J(w)A^\top\right)^{-1} J(w); \\ &\approx \frac{1}{\lambda} I_p \quad + \quad \frac{1}{2\lambda} A^\top \left(\lambda I_n - J(w)A^\top\right)^{-1} J(w) + \\ &\quad + \quad \frac{1}{2\lambda} J(w)^\top \left(\left(\lambda I_n - J(w)A^\top\right)^{-1}\right)^\top A.\end{aligned}$$

This method only requires the inversion of the $n \times n$ sparse matrix $\lambda I_n - J(w)A^\top$. Due to the sparsity of $J(w)$, the resulting matrix H^{-1} might be sparse too. If this is the case, the inverse of the Hessian matrix can be easily stored in memory.

A similar method can be used when the inverse of the Hessian matrix is part of an intermediate step for calculating the variance of the calibrated totals. Thus, the variance of the totals $\hat{y} = A\hat{w}$ can be approximated as

$$\mathsf{VAR}(\hat{y}) \approx P_1 + \frac{1}{2\lambda}(P_2 P_3^{-1} P_4 + (P_3^{-1} P_4)^\top P_2),$$

where the P_i are $n \times n$ matrices for any $i = 1, \ldots, 4$, and they are defined as

$$\begin{aligned} P_1 &= \frac{1}{\lambda} A D^2 A^\top, & P_2 &= A D A^\top, \\ P_3 &= \lambda I_n - J(w) A^\top, \quad \text{and} & P_4 &= J(w) D A^\top, \end{aligned}$$

where the diagonal matrix D is formulated to take into account possible adjustments due to any theoretical maximum variance allowed by the constraints in (2). In fact, when the i-th weight w_i is restricted to be within the interval $[\ell_w, u_w]$, the following inequality holds:

$$\mathsf{VAR}(\hat{w}_i) \leq \left(\frac{u_w - \ell_w}{2}\right)^2,$$

and hence the i-th diagonal entry of D is computed as

$$D_{ii} = \begin{cases} \dfrac{u_w - \ell_w}{2\left(H^{-1}\right)_{ii}^{0.5}}, & \text{if } \left(H^{-1}\right)_{ii} > \left(\dfrac{u_w - \ell_w}{2}\right)^2, \\ 1, & \text{otherwise.} \end{cases}$$

However, this requires the knowledge of the i-th diagonal entries of the matrix H^{-1}.

3. Bootstrap methods

Bootstrap methods are applied to make inference by simulating the distribution of a statistic of interest. Here they are used when the covariance matrices of the totals or weights are not positive-definite due to the approximations involved in the method presented in the previous section. If this is the case, the re-sampling techniques described below can be considered instead.

3.1 Parametric bootstrap

In this context, the idea behind the parametric bootstrap is mainly justified if the estimates are assumed to be normally distributed. Indeed, if $w \sim N_p(\hat{w}, \hat{\Sigma}_{\hat{w}})$, then $y \sim N_n(\hat{y}, \hat{\Sigma}_{\hat{y}})$, where the matrix $\hat{\Sigma}_{\hat{w}} \approx \left(\lambda I_p - A^\top J(\hat{w})\right)^{-1}$, and $\hat{\Sigma}_{\hat{y}} \approx P_1 + \frac{1}{2\lambda}(P_2 P_3^{-1} P_4 + (P_3^{-1} P_4)^\top P_2)$. This means that B random vectors must be generated from a multivariate normal distribution, and only after, the variance for each weight can be calculated as

$$\mathsf{VAR}(\hat{w}_i) = \frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{w}_i^{(b)} - \hat{w}_i\right)^2,$$

where $w_i^{(b)}$ is the i-th component of the b-th simulated vector of weights, and similarly,

$$\mathsf{VAR}(\hat{y}_i) = \frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{y}_i^{(b)} - \hat{y}_i\right)^2$$

for the totals. The covariance between two weights can be computed as

$$\mathsf{COV}(\hat{w}_i, \hat{w}_j) = \frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{w}_i^{(b)} - \hat{w}_i\right)\left(\hat{w}_j^{(b)} - \hat{w}_j\right),$$

and the covariance between two totals are obtained analogously.

The main issue of this technique is related to the simulation of a multivariate normal distributed vector whose covariance matrix Σ might be non-positive definite. This simulation requires the computation of $\Sigma^{1/2}$, which can be performed through the singular value decomposition (SVD) so that $\Sigma = U\Delta V^\top$, where U and V are two orthogonal matrices, and Δ is a diagonal one. In so doing, $\Sigma^{1/2} = U\Delta^{1/2}V^\top$, and therefore the vector of simulated weights can be computed as

$$w^{(b)} = \hat{w} + U\Delta^{1/2}V^\top x,$$

where x is a vector of independent random numbers drawn from a standard normal distribution. Similarly, the same computations are used for estimating the variances of the totals.

3.2 Nonparametric bootstrap

The nonparametric bootstrap approach is an appealing solution for the estimation of the variance of the estimated totals. In fact, without assuming the normality of the weights, it is still possible to perform a re-sampling scheme from the data collected. In this context, the standard *ad hoc* techniques (Beaumont and Patak, 2012) face conceptual contradictions with the calibration setup. However, these problems do not arise when the plug-in approach is adopted instead.

The rationale of the plug-in approach is based on the mimicking principle, where the bootstrap samples are selected from the estimated population (Ranalli and Mecatti, 2012). Therefore, the probability that the observation s_i from a finite population $\mathcal{U}$ belongs to the sample $\mathcal{S}_b$ is given by

$$\Pr(s_i \in \mathcal{S}_b) = \frac{1}{\hat{w}_i}, \tag{4}$$

for any $b = 1, \ldots, B$. This means that the observation s_i in the sample $\mathcal{S}_b$ can be replaced by other $\hat{w}_i$ equivalent units in the population (Antal and Tillé, 2011). According to this

principle, the algorithm estimates the variance of the totals by performing a randomized sampling with replacement form the entire population $\mathcal{U}$.

In order to avoid the construction of the entire estimated population, the units are sampled directly from those processed by the calibration algorithm. Each observation can be selected multiple times due to the chosen re-sampling scheme, so that the observation s_i will be randomly replicated $M_i^{(b)}$ times, where $M_i^{(b)} \sim \mathsf{Bin}\left(\hat{w}_i, \hat{w}_i^{-1}\right)$. The bootstrapped totals are computed afterward as

$$\hat{y}_j^{(b)} = \sum_{i=1}^{p} a_{ji}\hat{w}_i M_i^{(b)},$$

so that the expectation of $\hat{y}_j^{(b)}$ is

$$\mathsf{E}\left[\hat{y}_j^{(b)}\right] = \sum_{i=1}^{p} a_{ji}\hat{w}_i = \hat{y}_j,$$

for any $j = 1, \ldots, n$, since $\mathsf{E}\left[M_i^{(b)}\right] = 1$ for any $i = 1, \ldots, p$. The variance is then estimated as

$$\mathsf{VAR}(\hat{y}_j) = \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{y}_j^{(b)} - \bar{y}_j\right)^2, \tag{5}$$

and the coefficients of variation are calculated as

$$\mathsf{CV}(\hat{y}_j) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\frac{\hat{y}_j^{(b)}}{\bar{y}_j} - 1\right)^2},$$

for any $j = 1, \ldots, n$, where $\bar{y}_j$ denotes the arithmetic average over the bootstrapped totals $\hat{y}_j^{(b)}$, i.e.

$$\bar{y}_j = \frac{1}{B}\sum_{b=1}^{B}\hat{y}_j^{(b)}.$$

The estimator in (5) provides the exact Horvitz-Thompson estimator (Horvitz and Thompson, 1952) for the variance of the estimated totals (Antal and Tillé, 2011). This means that, under proper conditions, it is an unbiased estimator.

4. Simulation study

The variance estimation techniques presented above were compared on a framework where the data are collected from a simulated population. The data of 10000 units of the population were artificially generated for $n = 150$ point targets. Thus, a generic component of the sparse data-matrix A was obtained by the product of two random variables, i.e.

$$a_{ij} = G_{ij}D_{ij},$$

where $D_{ij} \sim \mathsf{Ber}(0.03)$, and $G_{ij} \sim \mathsf{Poi}(|\kappa_i|)$, for any $i = 1, \ldots, n$, and any $j = 1, \ldots, p$, where $\kappa_i \sim \mathsf{N}(\mu_i, \sigma_i)$, with $\mu_i \sim \mathsf{Gam}(20, 0.5)$ and $\sigma_i \sim \mathsf{InvGam}(0.5, 1)$. A design weight w_j^* was then assigned to each of the artificial units for $j = 1, \ldots, 10000$, so that the entire reference population for the study could be formed by considering w_j^* equivalent units in the population. These weights were simulated from the distribution exposed in Table 1,

Table 1: Distribution used to simulate the design weights.

N. of eq. units	1	2	3	4	5	6	7	8
Probability	0.64	0.34	0.015	0.003	0.001	0.0005	0.00035	0.00015

and used to compute the total units in the population as $N = \sum_{j=1}^{10000} w_j^* = 13848$, and the point targets as $y = Aw^*$.

In order to compare the proposed methodologies for the variance estimation, 1000 random sampling replications were performed by selecting $p = 10000$ units from the reference population previously generated. The design weights were then used as the initial values for the integer programming calibration algorithm (Sartore and Toppin, 2016). However, if the weight $w_j^* > 6$, then the initial value of $\hat{w}_j$ was forced to be 6. Once the calibration weights $\hat{w}$ were obtained, the variance for the calibrated totals were computed via linear approximation, parametric and non-parametric bootstrap. The standard deviations were successively calculated. The distributions obtained from these quantities were represented in Figure 1. From these graphics, it is evident how the linear approximation method provides precise estimates, while the non-parametric bootstrap is the most variable. The three distributions also appear to be centered on the same value. In order to assess the bias of these estimates, Figure 2 was produced by drawing the expected standard deviation computed with the three different approaches. Since all the points are lying along the 45° lines, one can conclude that the three proposed estimators are essentially unbiased, since the non-parametric estimator is unbiased. From a computational perspective, the parametric bootstrap is the most intensive algorithm, while the non-parametric bootstrap is the most efficient (see Figure 3).

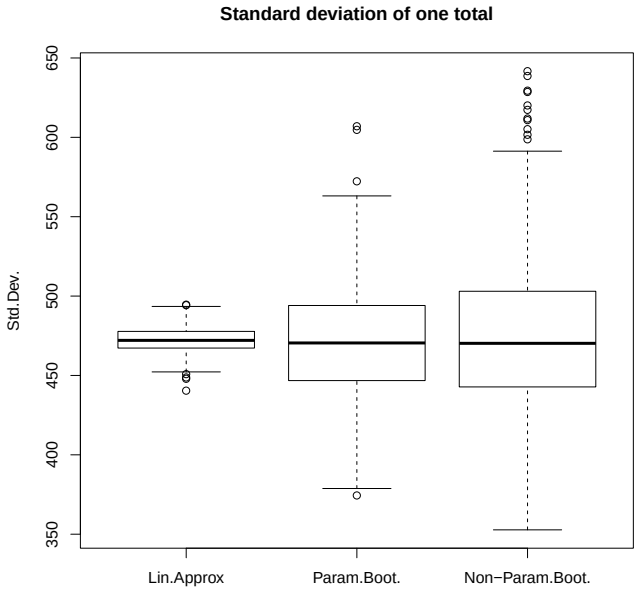

Figure 1: Distribution of Standard Deviations.

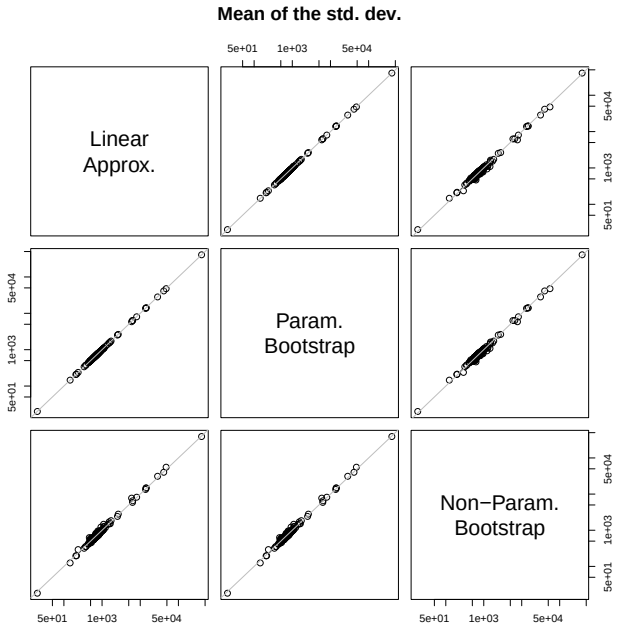

Figure 2: The average of the standard deviation for each total computed through each proposed algorithm.

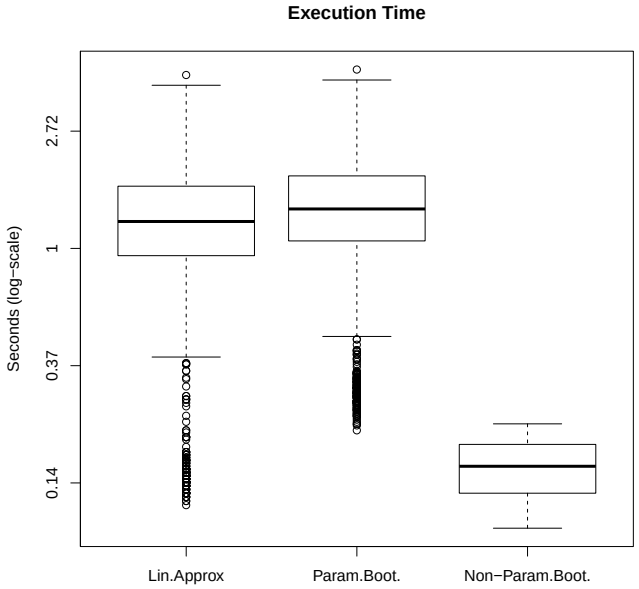

Figure 3: Computational efficiency.

5. Discussion

The three variance-estimation methods capture the uncertainty of the calibration process, and produce essentially unbiased estimates of the standard deviations while reducing the computational burden and memory usage. In particular, these methodologies allow for the estimation of the covariance matrix of totals without having the knowledge of the estimated

covariance matrix of weights.

The estimation of the covariance matrix of the weights can be computationally demanding when performing the parametric bootstrap. In fact, this technique needs to perform an SVD algorithm to compute the matrix $\Sigma^{1/2}$, and this can be intensive when p is large. This has two main computational issues. First of all, additional operations are required to simulate from the distribution of the weights, and second, it is not a guarantee that there is enough memory to store all the significant values of each matrix provided by the SVD algorithm. However, Baglama (2016) presented algorithms to obtain the SVD with sparse matrices such that the computational burden is less intensive than the standard algorithms.

Massively large covariance matrices are usually represented in sparse data structures, but these are not suitable when the matrices are dense. To obviate this problem a specific data structure can be designed to store both the standard deviations and the lower triangular matrix of the correlations in order to compute the covariances. In addition, a digital representation in 16-bit floating points allows for the whole structure to be stored into the computer's memory at a cost of losing some accuracy and efficiency of the estimates.

Such structures are not needed by the non-parametric bootstrap, because the covariance matrix of the weights is assumed to be diagonal with variances given by the selection probabilities in (4), i.e. $\mathsf{VAR}\left[w_i M_i | w_i\right] = w_i(w_i - 1)$. This assumption is equivalent to the pairwise-independence when the weights are assumed to be normally distributed. Unfortunately, biased estimates of the covariances can be produced by the non-parametric bootstrap. In addition, the independence of the weights is inconsistent when the totals are assumed to be independent during the calibration, even if calibration is performed simultaneously for all the targets. In fact, the relationship $\Sigma_y = A\Sigma_w A^\top$ is almost always false, when A is an $n \times p$ matrix and Σ_y and Σ_w are both diagonal matrices.

Further research is required to address the methodological consistency due to the independence assumptions. However, the methodologies presented in this article allow for the production of approximately unbiased variances for the totals. The outputs from these three methods can be used as solid guidelines in the calculation of a benchmark diagonal matrix Σ_y, which is used in the calibration of the covariance matrix of the weights. The diagonal entries of the matrix Σ_y are the unbiased variances of the totals. Therefore, a consistent optimization method needs to be developed to deal with such high-dimensional matrices, which demand extraordinary capabilities and new sophisticated techniques, while maintaining both the standard levels of accuracy and computational efficiency.

6. Conclusion

In this paper, we presented three different methods to estimate high-dimensional variance-covariance matrices associated with the calibration weights. These methods are based on classical approaches, where the linear approximation and the bootstrap are especially considered. In most cases, the estimation of a huge covariance matrix can be bypassed by computing directly the covariance matrix of the calibrated totals. The proposed estimators are shown by simulation to be unbiased, and each of them has its own strengths and weaknesses in terms of precision and computational efforts.

References

Antal, E. and Tillé, Y. (2011). A direct bootstrap method for complex sampling designs from a finite population. *Journal of the American Statistical Association*, 106(494):534–543.

Baglama, J. (2016). Irlba: Fast partial singular value decomposition method. *Handbook of Big Data*, page 125.

Beaumont, J.-F. and Patak, Z. (2012). On the generalized bootstrap for sample surveys with special attention to poisson sampling. *International Statistical Review*, 80(1):127–148.

Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87(418):376–382.

Fuller, W. A. (1975). Regression analysis for sample survey. *Sankhya, Series C*, 37:117–132.

Furno, M. (1998). Estimating the variance of the LAD regression coefficients. *Computational statistics & data analysis*, 27(1):11–26.

Henderson, H. V. and Searle, S. R. (1980). On deriving the inverse of a sum of matrices. *Biometrics Unit series*.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685.

Kott, P. S. (1999). Some problems and solutions with a delete-a-group jackknife. In *Federal Committee on Statistical Methodology Research Conference*, volume 4, pages 129–135.

Kott, P. S. (2016). Calibration weighting in survey sampling. *Wiley Interdisciplinary Reviews: Computational Statistics*, 8(1):39–53.

Lemel, Y. (1976). Une généralisation de la méthode du quotient pour le redressement des enquêtes par sondage. In *Annales de l'INSEE*, pages 273–282.

Ranalli, M. G. and Mecatti, F. (2012). Comparing recent approaches for bootstrapping sample survey data: A first step towards a unified approach. In *Proceedings of the Survey Research Methods Section of the American Statistical Association*, pages 4088–4099.

Särndal, C.-E., Swensson, B., and Wretman, J. H. (1989). The weighted residual technique for estimating the variance of the general regression estimator of the finite population total. *Biometrika*, 76(3):527–537.

Sartore, L. and Toppin, K. (2016). *inca: Integer Calibration*. R package version 0.0.2.

Singh, A. C. and Folsom, R. E. (2000). Bias corrected estimating function approach for variance estimation adjusted for poststratification. In *Proceedings of the Survey Research and Methodology Section of the American Statistical Association*, pages 610–615.

Valliant, R. (2007). An overview of the pros and cons of linearization versus replication in establishment surveys. In *Proceedings of the Third International Conference on Establishment Surveys*, pages 929–940.

Wolter, K. M. (1985). *Introduction to Variance Estimation*. Springer.

Woodbury, M. A. (1950). Inverting modified matrices. *Memorandum report*, 42:106.